The Annals of Applied Statistics
2010, Vol. 4, No. 4, 1660–1697
DOI: 10.1214/10-AOAS363

SUBSAMPLING METHODS FOR GENOMIC INFERENCE[1]

By Peter J. Bickel*, Nathan Boley*, James B. Brown*, Haiyan Huang* and Nancy R. Zhang*

University of California at Berkeley, University of California at Berkeley, University of California at Berkeley, University of California at Berkeley, and Stanford University

Large-scale statistical analysis of data sets associated with genome sequences plays an important role in modern biology. A key component of such statistical analyses is the computation of p-values and confidence bounds for statistics defined on the genome. Currently such computation is commonly achieved through ad hoc simulation measures. The method of randomization, which is at the heart of these simulation procedures, can significantly affect the resulting statistical conclusions. Most simulation schemes introduce a variety of hidden assumptions regarding the nature of the randomness in the data, resulting in a failure to capture biologically meaningful relationships. To address the need for a method of assessing the significance of observations within large scale genomic studies, where there often exists a complex dependency structure between observations, we propose a unified solution built upon a data subsampling approach. We propose a piecewise stationary model for genome sequences and show that the subsampling approach gives correct answers under this model. We illustrate the method on three simulation studies and two real data examples.

1. Introduction.

1.1. *Background.* This paper grew out of a number of examples arising in data coming from the Encyclopedia of DNA Elements (ENCODE) Pilot Project [Birney et al. (2007)], which is composed of multiple, diverse experiments performed on a targeted 1% of the human genome. Computational analyses of this data are aimed at revealing new insights about how

Received July 2009; revised March 2010.
[1]Supported by NIH U01 HG004695 (ENCODE DAC) and NIH 5R01GM075312.
*The authors are ordered alphabetically.
Key words and phrases. Genome Structure Correction (GSC), subsampling, piecewise stationary model, segmentation-block bootstrap, feature overlap.

the information coded in the DNA blueprint is turned into functioning systems in the living cell. Variations of some of the methods described here have been applied at various places in that paper, as well as in Margulies et al. (2007), for assessing significance and computing confidence bounds for statistics that operate along a genomic sequence. The background of these methods is described in cookbook form in the supplements to those papers, and it is the goal of this paper to present them rigorously and to develop some necessary refinements.

Essentially, we will argue that, in making inference about statistics computed from "large" stretches of the genome, in the absence of real knowledge about the evolutionary path which led to the genome in question, the best we can do is to model the genome by a piecewise stationary ergodic random process. The variables of this process can be base pair composition or some other local features, such as various annotated functional elements.

In the purely stationary case some of the types of questions that we will address, such as tests for independence of point processes, confidence bounds for expectations of local functions and goodness of fit of models, have been considered extensively. However, inference for piecewise stationary models appears not to have been investigated. With the advent of enormous amounts of genomic data, all sorts of inferential questions have arisen. The proposed model may be the only truly nonparametric approach to the genome, although, just as in ordinary nonparametric statistics, there are many possible ways of carrying out inference.

Our methods are based on a development of the resampling schemes of Politis and Romano (1994), Politis, Romano and Wolf (1999) and the block bootstrap methods of Künsch (1989). As we shall see, in many situations, Gaussian approximations can replace these schemes. But in these situations, as with the ordinary bootstrap, we believe that a subsampling approach is valuable for the following reasons:

- Letting the computer do the approximation is much easier.
- Some statistics, such as tests of the Kolmogorov–Smirnov type, are functions of stochastic processes to which a joint Gaussian approximation applies. Then, limiting distributions can only be computed by simulation.
- The bootstrap distributions of our statistics show us whether the approximate Gaussianity we have invoked for the "true" distribution of these statistics is in fact warranted. This visual confirmation is invaluable.

This paper is organized as follows. We begin with some concrete examples from the ENCODE data as well as other types of genomic data in Section 1.2, and proceed with a motivated description of our model in Section 2. Our methods are discussed both qualitatively and mathematically in Sections 3 and 4. Section 5 contains results from simulation studies and real data

analysis. Proofs of theorems stated in Sections 3 and 4 can be found in the supplemental article [Bickel et al. (2010)].

The statistics and methods discussed in this paper have been implemented in several computing languages and are available for download at http://www.encodestatistics.org/. Each of these implementations runs in $n\log(n)$ time, where n is the number of instances of the more frequent feature. On a desktop PC (Intel Core Duo 3 GHz and 2 Gb RAM) the Python version takes over 1000 samples per second for features on the order of 10^4 instances.

1.2. *Motivating examples.* We start with several fundamental questions that arise in genomic studies.

- *Association of functional elements in genomes.* In genomic analyses, a natural quantity of interest is the association among different functional sites/features annotated along the DNA sequence. Its biological motivation comes from the common belief that significant physical overlapping or proximity of functional sites in the genome suggests biological constraints or relationships. In the ENCODE project, to understand the possible functional roles of the evolutionarily constrained sequences that are conserved across multiple species, overlap between the constrained sequences and several experimental annotations, such as 5'UTR, RxFrags, pseudogenes, and coding sequences (CDSs), have been evaluated using the method discussed in this paper. It was found that the overlap of most experimental annotations with the constrained sequences are significantly different from random [Birney et al. (2007)]. An illustrative example from The ENCODE Project [Birney et al. (2007)] is detailed in Section 5.1.
- *Cooperativity between transcription factor binding sites.* In some situations, there is interest to study the associations between neighboring functional sites that do not necessarily overlap. For instance, it is known that transcription factors often work cooperatively and their binding sites (TFBS) tend to occur in clusters [Zhang et al. (2006)]. Consequently, methods for identifying interacting transcription factors usually involve evaluating the significance of co-occurrences of their binding sites in a local genomic region [Zhou and Wong (2004); Das, Banerjee and Zhang (2004); Yu, Yoo and Greenwald (2004); Huang et al. (2004); Kato et al. (2004); Gupta and Liu (2005)]. This problem has the same formulation as the above ENCODE examples given a functional site defined as follows: for a TFBS of length l at position i, we define the region $(i-m, i+l+m)$ as a functional site. Then two overlapping functional sites are equivalent to two neighboring TFBSs with interdistance less than $2m$, and the methods discussed in this paper for evaluating the significance of overlapping functional features can be applied. We leave this and related applications which involve considering statistics of the K-S type to a later paper.

- *Correlating DNA copy number with genomic content.* Recent technology has made it possible to assay DNA copy number variation at a very fine scale along the genome [for review, see Carter (2007)]. Many studies, for example, Redon et al. (2006), have shown that such variation in DNA copy number is a common type of polymorphism in the human genome. To what extent do these regions of copy number changes overlap with known genomic features, such as coding sequences? Redon et al. performed such an analysis and argued that copy number variant regions have a significant paucity for coding regions. The p-values supporting this claim were based on random permutations of the start locations of the variant segments. This assumes uniformity and stationarity of the copy number variants. However, CNVs do not occur at random and are often clustered in regions of the genome containing segmental duplications. The methods discussed in this paper for evaluating the significance of overlapping features, which assume neither uniformity nor stationarity, can again be applied to this problem. Actually, the results from our method suggest a different conclusion on this problem (see Section 5.5).

As we have seen in these examples, a common question asked in many applications is the following: Given the position vectors of two features in the genome, for example, "conservation between species" and "transcription start sites," and a measure of relatedness between features, for example, base or region percentage overlap, how significant is the observed value of the measure? How does it compare with that which might be observed "at random?"

The essential challenge in the statistical formulation of this problem is the appropriate modeling of randomness of the genome, since we observe only one of the multitudes of possible genomes that evolution might have produced for our and other species.

How have such questions been answered previously? Existing methods employ varied ways to simulate the locations of features within genomes, but all center around the uniformity assumption of the features' start positions: The features must occur homogeneously in the studied genome region, for example, Blakesley et al. (2004) and Redon et al. (2006). This assumption ignores the natural clumping of features as well as the nonstationarity of genome sequences. Clumping of features is quite common along the genome due to either the feature's own characteristic, for example, transcription factor binding sites (TFBSs) tend to occur in clusters, or the genome's evolutionary constraints, for example, conserved elements are often found in dense conservation neighborhoods. Ignoring these natural properties could result in misleading conclusions.

In this paper we suggest a piecewise stationary model for the genome (see Section 2) and, based on it, propose a method to infer the relationships

between features which we view as "nonparametric" as possible (see Sections 4.2 and 4.4). The model is based on assumptions which we demonstrate in real data examples to be plausible.

2. The piecewise stationary model.

2.1. *Genomic motivation.* We postulate the following for the observed genomes or genomic regions:

- They can be thought of as a concatenation of a number of regions, each of which is homogenous in a way we describe below.
- Features that are located very far from each other on the average have little to do with each other.
- The number of such homogeneous regions is small compared to the total length of the observed genome.

These assumptions, which form the underpinning of our *block stationary model* for genomic features, are motivated by earlier studies of DNA sequences, which show that there are global shifts in base composition, but that certain sequence characteristics are locally unchanging. One such characteristic is the GC content. Bernardi et al. (1985) coined the term "isochore" to denote large segments (of length greater than 300 Kb) that have fairly homogeneous base composition and, especially, constant GC composition. Even earlier, evidence of segmental DNA structure can be found in chromosomal banding in polytene chromosomes in drosophila, visible through the microscope, that result from underlying physical and chemical structure. These banding patterns are stable enough to be used for the identification of chromosomes and for genetic mapping, and are physical evidence for a block stationarity model for the GC content of the genome.

The experimental evidence for segmental genome structure and the increasing availability of DNA sequence data have inspired attempts to computationally segment DNA into statistically homogeneous regions. The paper by Braun and Müller (1998) offers a review of statistical methods developed for detecting and modeling the inhomogeneity in DNA sequences. There have been many attempts to segment DNA sequences by both base composition [Fu and Curnow (1990); Churchill (1989, 1992); Li et al. (2002)] and chemical characteristics [Li et al. (1998)]. Most of these computational studies concluded that a model that assumes block-wise stationarity gives a significantly better fit to the data than stationary models [see, for example, the conclusions of two very different studies by Fickett, Torney and Wolf (1992) and Li et al. (1998)].

A subtle issue in the definition of "homogeneity" is the scale at which the genome is being analyzed. Inhomogeneity at the kilobase resolution, for example, might be "smoothed out" in an analysis at the megabase level.

The level of resolution is a modeling issue that must be considered carefully with the goal of the analysis in mind.

Implicit in our formulation is an "ergodic" hypothesis. We want probabilities to refer to the population of potential genomes. We assume that the statistics of the genome we have mimic those of the population of genomes. This is entirely analogous to the ergodic hypothesis that long term time averages agree with space averages for trajectories of dynamic systems.

2.2. *Mathematical formulation.* In mathematical terms, the block stationarity model assumes that we observe a sequence of random variables $\{X_1, \ldots, X_n\}$ positioned linearly along the genomic region of interest. $X_k, k = 1, \ldots, n$, may be base composition, or some other measurable feature. We assume that there exist integers $\boldsymbol{\tau} = \boldsymbol{\tau}^{(n)} = (\tau_0, \ldots, \tau_U)$, where $0 = \tau_0 < \tau_1 < \cdots < \tau_U = n$, such that the collections of variables, $\{X_{\tau_i}, \ldots, X_{\tau_{i+1}}\}$, are separately stationary for each $i = 0, \ldots, U-1$. We let $n_i = \tau_i - \tau_{i-1}$ be the length of the ith region, and let there be U such regions in total. For convenience, we introduce the mapping

$$\pi : \{1, \ldots, n\} \to \{(i, j) : 1 \le i \le U, 1 \le j \le n_i\}$$

which relates the relabeled sequence, $\{X_{ij} : 1 \le i \le U, 1 \le j \le n_i\}$, to the original sequence $\{X_1, \ldots, X_n\}$. We write $\pi = (\pi_1, \pi_2)$ with $\pi(k) = (i, j)$ if and only if $k = \tau_i + j$. We will use the notation X_{ij} and X_k interchangeably throughout this paper.

For any k_1, k_2, let $\mathcal{F}_{k_1}^{k_2}$ be the σ-field generated by $X_{k_1}, \ldots, X_{k_2}$. Define $m(k)$ to be the standard Rosenblatt mixing number [Dedecker et al. (2007)],

$$m(k) = \sup\{|\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{F}_1^l, B \in \mathcal{F}_{l+k}^n, 1 \le l \le n - k\}.$$

Then, assumptions 1–3 stated in Section 2.1 translate to the following:

A1. The sequence $\{X_1, \ldots, X_n\}$ is piecewise stationary. That is, $\{X_{ij} : 1 \le j \le n_i\}$ is a stationary sequence for $i = 1, \ldots, U$.
A2. There exists constants c and $\beta > 0$ such that $m(k) \le ck^{-\beta}$ for all k.
A3. $U/n \to 0$.

An immediate and important consequence of A1–A3 is that, for any fixed small k, if we define $W_1 = (X_1, \ldots, X_k), W_2 = (X_{k+1}, \ldots, X_{2k}), \ldots, W_m = (X_{n-k+1}, \ldots, X_n)$, where $m = n/k$, then $\{W_1, \ldots, W_m\}$ also obey A1–A3. This is useful, for example, in the region overlap example considered in the next section.

The remarkable feature of these assumptions, which are more general to our knowledge than any made heretofore in this context, is that they still allow us to conduct most of the statistical inference of interest. Not surprisingly, these assumptions lead to more conservative estimates of significance than any of the previous methods.

3. Vector linear statistics and Gaussian approximation. We study the distribution of a class of vector linear statistics of interest under the above piecewise stationary model. As an illustration, we consider the ENCODE data examples, and suppose that we are interested in base pair overlap between features A and B. We can represent base pair overlap by defining

$$I_k = 1, \qquad \text{if position } k \text{ belongs to feature } A \text{ and 0 otherwise,}$$
$$J_k = 1, \qquad \text{if position } k \text{ belongs to feature } B \text{ and 0 otherwise.}$$

We can then define $X_k = I_k J_k$ to be the indicator that position k belongs to both features A and B. Then, for the $n = 30$ megabases of the ENCODE regions, the mean base pair overlap is equal to

$$\bar{X} = \sum_{k=1}^{n} X_k / n.$$

Another biologically interesting statistic is the (asymmetric) region overlap, defined as follows: suppose the consecutive feature stretches are $T_1, \ldots, T_\alpha$ with lengths $\tau_1, \ldots, \tau_\alpha$, and the corresponding nonfeature stretches $S_1, \ldots, S_\beta$ with lengths $\rho_1, \ldots, \rho_\beta$. We assume here that the initial and final stretches consist of one feature and one nonfeature stretch. The complementary situation, when both initial and final stretches are of the same type, is dealt with similarly. Without loss of generality, suppose the initial stretch is nonfeature. Then, $S_1 = \{1, \ldots, \rho_1\}$, $T_1 = \{\rho_1 + 1, \ldots, \rho_1 + \tau_1\}$, $S_2 = \{\rho_1 + \tau_1 + 1, \ldots, \rho_1 + \tau_1 + \rho_2\}$, etc. Using I_k, J_k as indicators of feature identity, we define the (unnormalized) region overlap of feature A stretches with feature B stretches as $\frac{1}{n}\sum_{t=1}^{\alpha} V_t$ where $V_t = 1 - \prod_{k \in T_{A,t}} (1 - J_k)$, where $T_{A,1}, \ldots, T_{A,\alpha}$ denote the feature A stretches. This statistic is not linear in terms of functions of single basepairs, but is linear in functions of blocks of feature B. These blocks are of random sizes, but are consistent with our hypothesis of piecewise stationarity that, except for end effects due to feature instances crossing segment boundaries, the V_t are also stationary. If the lengths $\tau_1, \ldots, \tau_\alpha$ are negligible compared to n and α is of the order of n, we can expect the mixing hypothesis to remain valid.

In general, we focus our attention on statistics that can be expressed as a function of the mean of $\mathbf{g}(X_i)$, where $\mathbf{g}$ is some well behaved d-dimensional vector function to be characterized in later sections. By the flexible definition of $\mathbf{g}$, this encompasses a wide class of situations.

First, we consider vector linear statistics of the form

$$\mathbf{T}_n(\mathbf{X}) = n^{-1} \sum_{k=1}^{n} \mathbf{g}(X_k).$$

We introduce the following notation:

$$E[\mathbf{T}_n] \equiv \boldsymbol{\mu} \equiv \sum_{i=1}^{U} f_i \boldsymbol{\mu}_i,$$

where

$$\boldsymbol{\mu}_i \equiv E[\mathbf{g}(X_{i1})],$$
$$f_i \equiv n_i/n$$

and

$$\Sigma_n \equiv \mathbb{V}\mathrm{ar}(n^{1/2}\mathbf{T}_n) = \sum_{i=1}^{U} f_i C_i(nf_i), \tag{3.1}$$

where

$$C_i(m) = C_{i0} + 2\sum_{\ell=1}^{m} C_{i\ell}\left(1 - \frac{(\ell-1)}{m}\right)$$

and

$$C_{i0} \equiv \mathbb{V}\mathrm{ar}\,\mathbf{g}(X_1), \qquad C_{i\ell} \equiv \mathbb{C}\mathrm{ov}(\mathbf{g}(X_{i1}), \mathbf{g}(X_{i(l+1)})). \tag{3.2}$$

In Theorem 3.1 below, we show asymptotic Gaussianity of $\mathbf{T}_n$ given a few more technical assumptions:

A4. $\frac{1}{n}\sum_{i:n_i \le l} n_i \to 0$ for all $l < \infty$.
A5. $\forall i$, $|\mathbf{g}|_\infty \le C < \infty$.
A6. $0 < \varepsilon_0 \le \|\Sigma_n\| \le \varepsilon_0^{-1}$, for all n, where $\|\cdot\|$ is a matrix norm.

In particular, A4 implies that the contribution of "small regions" to the overall statistic must not be too large.

THEOREM 3.1. *Under conditions* A1–A6,

$$n^{1/2}\Sigma_n^{-1/2}(\mathbf{T}_n - \boldsymbol{\mu}) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3.3}$$

where $\mathbf{I}$ *is the* $d \times d$ *identity.*

The proof of the theorem is in the supplemental article [Bickel et al. (2010)]. If we have estimates $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ which are consistent in a suitably uniform sense, then estimates of $C_{i\ell}$, $C_i(m)$ using $\hat{\boldsymbol{\tau}}$ in place of $\boldsymbol{\tau}$ are also consistent. However, simply plugging these estimates into (3.1) does not yield consistent estimates of σ^2 if our approach were to compute confidence intervals by Gaussian approximation. This is well known for the stationary case. Some regularization is necessary. We do not pursue this direction but prefer to

approach the inference problem from a resampling point of view—see next section.

In many cases, the statistics of interest are not linear. For example, in the analysis of the ENCODE data a more informative statistic is the %bp overlap defined as

$$B \equiv \frac{\bar{X}}{D}, \tag{3.4}$$

where

$$D = \sum_{k=1}^{n} I_k$$

is the total base count of feature A.

More conceptually, the region overlap is

$$R \equiv \frac{1}{W_I} \sum_{k=1}^{K} V_k, \tag{3.5}$$

where $W_I = \sum_{k=1}^{n} I_{k-1}(1 - I_k)$, the number of feature A instances.

A standard delta method computation shows that the standard error of B can be approximated as follows: Let $\mu(D)$ and $\mu(\bar{X})$ be respectively the expectation of D and $\bar{X}$. Then,

$$\frac{\bar{X}}{D} - \frac{\mu(\bar{X})}{\mu(D)} \approx \frac{\bar{X} - \mu(\bar{X})}{\mu(D)} - \mu(\bar{X}) \frac{(D - \mu(D))}{\mu^2(D)},$$

and, hence, we can approximate $\frac{\bar{X}}{D}$ by a Gaussian variable with mean $\frac{\mu(\bar{X})}{\mu(D)}$ and variance

$$\sigma^2(B) \approx \frac{\sigma^2(\bar{X})}{\mu^2(D)} + \frac{\mu^2(\bar{X})}{\mu^4(D)} \sigma^2(D) - 2 \frac{\mu(\bar{X})}{\mu^3(D)} \operatorname{Cov}(\bar{X}, D), \tag{3.6}$$

where $\sigma^2(B)$, $\sigma^2(\bar{X})$, $\sigma^2(D)$ are the corresponding variances and $\operatorname{Cov}(\bar{X}, D)$ denotes the covariance. In doing inference, we can use the approximate Gaussianity of B with $\sigma^2(B)$ estimated using the above formula with regularized sample moments replacing the true moments.

We also note that goodness of fit or equality of population test statistics, such as Kolmogorov–Smirnov and many others, can be viewed as functions of empirical distributions, which themselves are infinite-dimensional linear statistics, and we expect, but have not proved, that the methods discussed in this paper and the underlying theories apply to those cases as well, under suitable assumptions.

4. Subsampling based methods. Here we propose a subsampling based approach, in particular, a combined segmentation-block subsampling method to conduct statistical inference under the piecewise stationary model, which we call "segmented block subsampling." In our method, the segmentation parameters governing scale are chosen first and then the size of the subsample is chosen based on stability criteria. The segmentation procedure, as we discussed, is motivated by the heterogeneity of large-scale genomic sequences. The block subsampling approach takes into account the local genomic structure, such as natural clumping of features, when conducting statistical inference. We explicitly demonstrate the advantages of using segmentation and block subsampling by simulation studies in Section 5.

4.1. *Stationary block subsampling.* Below we first review the results related to the stationary block bootstrap method in a homogeneous region ($U = 1$), and then show how the method breaks down when it is applied to a piecewise stationary sequence ($U > 1$).

4.1.1. *Review of results for the case of* $U = 1$. For completeness, we recall the following basic algorithm of Politis and Romano (1994) to obtain an estimate of the distribution of the statistic $\mathbf{T}_n(X_1, \ldots, X_n)$ under the assumption that the sequence $X_1, \ldots, X_n$ is stationary (i.e., $U = 1$).

ALGORITHM 4.1. (a) Given $L \ll n$ choose a number N uniformly at random from $\{1, \ldots, n - L\}$.

(b) Given the statistic $\mathbf{T}$, as above, compute

$$\mathbf{T}_L(X_{N+1}, \ldots, X_{N+L}) \equiv \mathbf{T}^*_{L1}.$$

(c) Repeat B times with replacement to obtain $\mathbf{T}^*_{L1}, \ldots, \mathbf{T}^*_{LB}$.

(d) Estimate the distribution of $\sqrt{n}(\mathbf{T}_n - \mu)$ by the empirical distribution $\mathcal{L}^*_B$ of

$$\left\{ \sqrt{\frac{n}{L}} [\mathbf{T}^*_{Lj} - \mathbf{T}_n(X_1, \ldots, X_n)], 1 \le b \le B \right\}.$$

By Theorem 4.2.1 of Politis, Romano and Wolf (1999),

$$\mathcal{L}^*_B \Longrightarrow \mathcal{N}_d(\mathbf{0}, \Sigma) \tag{4.1}$$

in probability for some constant Σ if (3.3) holds and if $L \to \infty$, $L/n \to 0$. As usual, convergence of $\mathcal{L}^*_B$ in law in probability simply means that if ρ is any metric for weak convergence on R^d, then $\rho(\mathcal{L}^*_B, \mathcal{L}) \xrightarrow{P} 0$.

Since all variables we deal with are in L_2, we take ρ to be the Mallows metric,

$$\rho^2_M(F, G) = \min\{E_P(W - V)^2 : P \text{ such that } W \sim F, V \sim G\}.$$

Useful properties of ρ_M are as follows:

(a) $\rho_M^2(\Sigma\pi_i F_i, \Sigma\pi_i G_i) \le \Sigma\pi_i \rho_M^2(F_i, G_i)$ for all $\pi_i \ge 0, \Sigma\pi_i = 1$ and
(b) If $F = F_1 * \cdots * F_m$, $G = G_1 * \cdots * G_m$, that is, F and G are distributions of sums of m independent variables, then $\rho_M^2(F, G) \le \sum_{i=1}^m \rho_M^2(F_i, G_i)$.

For convenience, when no confusion is possible, we will write $\rho_M(W, V)$ for $\rho_M(F, G)$ for random variables $W \sim F$, $V \sim G$.

4.1.2. *Performance of the block subsampling method in the piecewise stationary model when $U > 1$.* We turn to the analogue of Theorem 4.2.1 in Politis, Romano and Wolf (1999) for $U > 1$. We consider a vector linear statistic, for which the true distribution was described in Section 3. Here, we ask how Algorithm 4.1, which assumes stationarity, performs in this nonstationary context. We show that, in general, it does not give correct confidence bounds but is conservative, sometimes exceedingly so. The results depend on L, the subsample size, which is a crucial parameter in Algorithm 4.1. We sketch these issues in Theorem 4.2 below, for simplicity, letting g be the one-dimensional identity function $g(x) = x$. Let

$$\tau^2 = U^{-1} \sum_{i=1}^{U} (\mu_i - \bar{\mu})^2,$$

$$\bar{X}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} X_{ij}, \qquad \bar{X} \equiv n^{-1} \sum_{k=1}^{n} X_k = \sum_{i=1}^{U} f_i \bar{X}_i.$$

Also let

$$n_i^* \equiv \text{ Cardinality of } S_i \equiv \{k : k \in [N, N+L], \pi_1(k) = i\}$$

and

$$\bar{X}_i^* = 1(n_i^* \ne 0) \sum_j \{X_{ij} : j \in S_i\}/n_i^*,$$

$$\bar{X}_L^* = \sum_{i=1}^{U} f_i^* \bar{X}_i^* \qquad \text{where } f_i^* \equiv \frac{n_i^*}{L}.$$

We introduce one assumption that is obviously needed for any analysis of the block or segmented resampling bootstraps:

A7. $L \to \infty$,

and two other assumptions which are used in different parts of Theorem 4.2 but not in the rest of the paper, and are thus given a different numbering:

B1. $L/n \to 0$.
B2. $(LU)/n \to 0$.

THEOREM 4.2. *Let $\mathcal{L}_n$ be the distribution which assigns mass f_i to $(\mu_i - \mu)$, $1 \le i \le U$, and write C_i for $C_i(nf_i)$. Suppose assumptions* A1–A5 *and* A7 *hold:*

(i) *If* B2 *holds, $\rho_M(\bar{X}_L^* - \bar{X}, \mathcal{L}_n) \xrightarrow{P} 0$.*
(ii) *If*

$$\sum_{i=1}^{U} f_i(\mu_i - \mu)^2 = o(L^{-1}) \tag{4.2}$$

and B1 *holds, then*

$$\rho_M\left[\sqrt{L}(\bar{X}_L^* - \bar{X}), \sum_{i=1}^{U} f_i \mathcal{N}(0, C_i)\right] \xrightarrow{P} 0.$$

(iii) *If (4.2) and* B1 *hold and*

$$\sum_{i=1}^{U} f_i 1(|\Sigma_n - C_i| \ge \varepsilon) \to 0 \tag{4.3}$$

for all $\varepsilon > 0$, then

$$\rho_M(\sqrt{L}(\bar{X}_L^* - \bar{X}), \mathcal{N}(0, \Sigma_n)) \xrightarrow{P} 0.$$

The implications of Theorem 4.2 are as follows. If equation (4.2) does not hold, then $\bar{X}_L^* - \bar{X}$ does not converge in law at scale $L^{-1/2}$ so that it does not reflect the behavior of $L^{1/2}(\bar{X}_L - \mu)$ at all. This is a consequence of inhomogeneity of the segment means. Evidently in this case, confidence intervals of the percentile type for μ, $[\bar{X} + c_n(\alpha), \bar{X} + c_n(1-\alpha)]$, where $c_n(\alpha)$ is the α quantile of the distribution of $\bar{X}_L^* - \bar{X}$, will have coverage probability tending to 1, since $c_n(\alpha)$ and $c_n(1-\alpha)$ do not converge to 0 at rate $L^{-1/2}$ as they should. If B2 does not hold, we have to consider the possibility that $[N, N+L]$ covers K_N consecutive segments, whose total length is $o(n)$, such that the average over all such blocks is close to μ. However, in the absence of a condition such as (4.2) or mutual cancellation of μ_i^*, the scale of $\bar{X}_L^*$ will be larger than $L^{-1/2}$. These issues will be clarified by the proof of Theorem 4.2 in the supplemental article [Bickel et al. (2010)]. We note also that (4.2) can be weakened to requiring that the mean of blocks of consecutive segments whose total length is small compared to n be close to μ to order $o(L^{-1/2})$. But our statement makes the issues clear. Finally, note that B2 holds automatically if the number of segments U is bounded and if B1 holds.

If (4.2) does hold but (4.3) does not, then $\sqrt{L}(\bar{X}_L^* - \bar{X})$ converges in law to the Gaussian mixture $\sum_{i=1}^{U} f_i \mathcal{N}(0, C_i)$. The mixture of Gaussians is more

dispersed in a rough sense than a Gaussian with the same variance, which is

$$\sigma_n^2 = \sum_{i=1}^{U} f_i C_i;$$

see Andrews and Mallows (1974). Especially note that, if W has the mixture distribution and V is the Gaussian variable with the same variance, then

$$Ee^{tW} = \sum f_i e^{-(t^2/2)c_i} \geq e^{-t^2/2 \sum f_i C_i} = Ee^{tV}$$

by Jensen's inequality. This suggests that the tail probabilities will also be overestimated. The overdispersion here, which leads to conservativeness that is not as extreme as in case (i), is due to inequality of the variances from segment to segment. Finally, if (4.3) holds, then the segments have essentially the same mean and variance and stationary block subsampling works.

A mark of either (4.2) or (4.3) failing is a lack of Gaussianity in the distribution of $\bar{X}_L^* - \bar{X}$. This was in fact observed at some scales in the ENCODE project, which led us to crudely segment on biological grounds with reasonable success. However, the correct solution, which we now present in this paper, is to estimate the segmentation and appropriately adjust the subsampling procedure.

4.2. *A segmentation based block subsampling method.* We saw in the previous section that the naïve block subsampling method that was designed for the stationary case breaks down when the sequence follows a piecewise stationary model. We propose a stratified block subsampling strategy, which stratifies the subsample based on a "good" segmentation of the sequence which is estimated from the data. We first state the block subsampling method, and then in Section 4.2.3 give minimal conditions on the estimated segmentation for its consistency. In Section 4.3 we discuss possible segmentation methods.

4.2.1. *Description of algorithm.* Assume that we are given a segmentation $\mathbf{t} = (0 = t_0, t_1, \ldots, t_{m+1} = n)$, where m is the number of regions in $\mathbf{t}$. Assume that the total size L of the subsample is pre-chosen. We define a stratified block subsampling scheme as follows.

ALGORITHM 4.3. For $i = 1, \ldots, m$, let $\lambda_i = \lambda_i(t) = \lceil (t_i - t_{i-1})L/n \rceil$. We use the notation $X_{i;l}$ to denote the block of length l starting at i:

$$X_{i;l} = (X_i, \ldots, X_{i+l-1}).$$

Then, for each subsample,

Draw integers $\mathbf{N} = \{N_1, \ldots, N_m\}$, with N_i chosen uniformly from $\{t_{i-1} + 1, \ldots, t_i - \lambda_i(t) + 1\}$, and let

$$X^* = (X_1^*, \ldots, X_m^*) = (X_{N_1;\lambda_1(t)}, \ldots, X_{N_m;\lambda_m(t)}).$$

Repeat the above B times to obtain B subsamples: $X^{*,1}, \ldots, X^{*,B}$.

To obtain a confidence interval for $\boldsymbol{\mu}$, we assume that the statistic $\mathbf{T}_n$ has approximately a $N(\boldsymbol{\mu}, \Sigma_n/n)$ distribution as in the previous section. For each subsample drawn as described in Algorithm 4.3, compute the statistic $\mathbf{T}_L^{*,b} = \mathbf{T}_L^{*,b}(\mathbf{t}) = \mathbf{T}_L(X^{*,b})$. Form the sampling estimate of variance,

$$\widehat{\Sigma}_n \equiv \frac{L}{B} \sum_{b=1}^{B} (\mathbf{T}_L^{*,b} - \bar{\mathbf{T}}_L^*)'(\mathbf{T}_L^{*,b} - \bar{\mathbf{T}}_L^*), \tag{4.4}$$

where $\bar{\mathbf{T}}_L^* \equiv \sum_{b=1}^{B} \mathbf{T}_L^{*,b}/B$. We can now proceed to estimate the confidence interval for $\mathbf{T}_n$ in standard ways. For example, in the univariate case where $\sigma_n^2 \equiv \Sigma_n$ is a scalar:

(a) Use $\bar{X} \pm z_{1-\alpha/2} \frac{\widehat{\sigma}_n}{\sqrt{n}}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$th quantile of $N(0,1)$, for a $1-\alpha$ confidence interval.

(b) Efron's percentile method: Let $\bar{X}_{(1)}^* < \cdots < \bar{X}_{(B)}^*$ be the ordered $\bar{X}^{*,b}$, then use $[\bar{X}_{([B\alpha/2])}^*, \bar{X}_{([B(1-\alpha/2)])}^*]$ as a $1-\alpha$ confidence interval.

(c) Use a Studentized interval [Efron (1981)] or Efron's (1987) *BCA* method; see Hall (1992) for an extensive discussion.

Although the theory for (c) giving the best coverage approximation has not been written down, as it has been for the ordinary bootstrap, we expect it to continue to hold. Evidently, these approaches can be applied not only to vector linear statistics like $\mathbf{T}_n$ but also to smooth functions of vector linear statistics.

This algorithm assumes a given segmentation $\mathbf{t}$, which should be set to some good estimate $\hat{\boldsymbol{\tau}}^{(n)} = \{0 = \hat{\tau}_0, \hat{\tau}_1, \ldots, \hat{\tau}_m = n\}$ of the true change points $\boldsymbol{\tau}^{(n)}$. In order for the algorithm to perform well, a good segmentation is critical unless the sequence is already reasonably homogeneous. In Section 4.2.2 below we state the result that the algorithm is consistent if the given segmentation equals the true changepoints. Then, in Section 4.2.3, we state a few assumptions on the data determined segmentation $\hat{\boldsymbol{\tau}}^{(n)}$ which would enable us to act as if the segmentation were known and state Theorem 4.5 to that effect.

4.2.2. *Consistency with true segmentation.* Under the hypothetical situation where the segmentation $\mathbf{t}$ assumed in Algorithm 4.3 is equal to the

set of true changepoints, then the algorithm can be easily shown to be consistent. Here we state the result, which will be proved in the supplemental article [Bickel et al. (2010)].

First, we state a stronger version of the assumption B1, which requires that the square of the subsample size $L = L_n$ be $o(n)$:

A8. $L_n^2/n \to 0$.

Then, the consistency of Algorithm 4.3 given the true segmentation follows from the following theorem.

THEOREM 4.4. *If assumptions* A1–A8 *hold, then*

$$L_n^{1/2}{\Sigma_n}^{-1/2}[T^*_{L_n}(\tau_n) - T_n] \Rightarrow N(0, I) \tag{4.5}$$

in probability, where I is the $d \times d$ identity.

4.2.3. *Consistency with estimated segmentation.* Let

$$\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}^{(n)} = (\hat{\tau}_1^{(n)}, \ldots, \hat{\tau}_{\hat{U}_n}^{(n)})$$

be a segmentation of the sequence $X_1, \ldots, X_n$, which is determined from the data, and which is intended to estimate the true changepoints $\boldsymbol{\tau} = \boldsymbol{\tau}^{(n)}$. We will state conditions on $\hat{\boldsymbol{\tau}}$ such that the statistic obtained from Algorithm 4.3 based on $\hat{\boldsymbol{\tau}}$ is close to the statistic obtained from the same algorithm based on the true segmentation $\boldsymbol{\tau}$. This can be stated formally as follows. For any segmentation $\mathbf{t}$, let $\mathbf{X}^*(\mathbf{t})$ be a subsample drawn according to Algorithm 4.3 based on $\mathbf{t}$. Let $F^*_{n,\mathbf{t}}(\cdot)$ be the distribution of $\sqrt{L}\{T[\mathbf{X}^*(\mathbf{t})] - E^*T[\mathbf{X}^*(\mathbf{t})]\}$ conditioned on $X_1, \ldots, X_n$ and $\mathbf{t}$. Then, the desired property on the estimated segmentation $\hat{\boldsymbol{\tau}}$ is that

$$\rho_M^2[F^*_{n,\hat{\boldsymbol{\tau}}^{(n)}}, F^*_{n,\boldsymbol{\tau}^{(n)}}] \to_p 0, \qquad \text{as } n \to \infty, \tag{4.6}$$

where ρ_M^2 is the Mallows' metric described in Section 4.1.1. That is, for inferential purposes, $T[\mathbf{X}^*(\hat{\boldsymbol{\tau}})]$ has approximately the same distribution as $T[\mathbf{X}^*(\boldsymbol{\tau})]$. Then, since we have shown in Section 4.2.2 that

$$\rho_M^2[F^*_{n,\boldsymbol{\tau}^{(n)}}, \Phi(\Sigma_n)] \to_p 0,$$

where $\Phi(\Sigma_n)$ is the Gaussian distribution with mean 0 and variance Σ_n, (4.6) implies that

$$\sqrt{L_n}\Sigma_n^{-1}\{T[\mathbf{X}^*(\hat{\boldsymbol{\tau}}^{(n)})] - E^*T[\mathbf{X}^*(\mathbf{t})]\} \to N(0, I).$$

Let $\hat{n}_i = \hat{\tau}_{i+1}^{(n)} - \hat{\tau}_i^{(n)}$. We now state conditions on the estimated segmentation which guarantee (4.6):

A9. $\hat{U}_n/n \to 0$,

A10. $n^{-1}\sum_{i:\hat{n}_i\le k}\hat{n}_i \to 0$ for all $k<\infty$,

A11. $L_n n^{-1}\sum_{i=1}^{\hat{U}_n}\min_{1\le j\le\hat{U}_n}|\tau_i-\hat{\tau}_j| \to_p 0$.

Assumptions A9 and A10 for $\hat{\boldsymbol{\tau}}^{(n)}$ mirror assumptions A3 and A4 for $\boldsymbol{\tau}^{(n)}$. Assumption A11 is a consistency criterion: As the data set grows, the total discrepancy in the estimation of $\boldsymbol{\tau}^{(n)}$ by $\hat{\boldsymbol{\tau}}^{(n)}$ must be small.

THEOREM 4.5. *Under assumptions* A1–A11, (4.6) *holds.*

The proof is given in the supplemental article [Bickel et al. (2010)]. There are trivial extensions of this theorem to smooth functions of vector means, which are, in fact, needed but simply cloud the exposition.

Theorem 4.5 implies that confidence intervals based on subsamples

$$\{\mathbf{X}^{*,j}(\hat{\boldsymbol{\tau}}^{(n)}) : j=1,\dots,B\}$$

constructed by Algorithm 4.3 conditional on $\hat{\boldsymbol{\tau}}^{(n)}$ are consistent, as long as $\hat{\boldsymbol{\tau}}^{(n)}$ satisfies A9–A11. Here is the formal statement of this fact in the one-dimensional case, where $\hat{\sigma}_n^2$ replaces $\hat{\Sigma}_n$ and $\mathbf{g}$ is the identity.

COROLLARY 4.6. *Under assumptions* A1–A11*:*

1. *Let $\hat{\sigma}_n^2$ be the block subsampling estimate of variance defined in* (4.4)*, then*
$$P(\bar{X}-z_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n}<\mu<\bar{X}+z_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n}) \to_p 1-\alpha.$$
2. *Confidence intervals estimated by Efron's percentile method are consistent. That is,*
$$P(\bar{X}^*_{([n\alpha/2])}<\mu<\bar{X}^*_{([n(1-\alpha/2)])}) \to_p 1-\alpha.$$

4.3. *Segmentation methods.* The objective of the segmentation step is to divide the original data sequence $X_1,\dots,X_n$ into approximately homogeneous regions so that the variance estimated in Algorithm 4.3 approximates the true variance of T_n. A segmentation into regions of constant mean is sufficient for guaranteeing that Algorithm 4.3 gives consistent variance estimates. Therefore, we focus here on the segmentation of X into regions of constant mean.

In our simulation and data analysis, we use the dyadic segmentation approach, which we motivate and describe here using the simple case where g is the identity function. First consider the simple case where $X_1,\dots,X_n$ are independent with variance 1. In testing the null hypothesis

$$H_0 : E[X_i]=\mu,$$

versus the alternative H_A that there exists $1 < \tau < n$ such that $E[X_i] = \mu_1$ for $i < \tau$ and $E[X_i] = \mu_2 \neq \mu_1$ for $i \geq \tau$, one can show that the following is the generalized likelihood ratio test:

$$\text{Reject } H_0 \text{ if } \max_{1<j<n} nM(j) > c,$$

where

$$M(j) = \frac{j}{n}(\bar{X}_{1:j} - \bar{X}_{1:n})^2 + \frac{n-j}{n}(\bar{X}_{j+1:n} - \bar{X}_{1:n})^2. \tag{4.7}$$

The maximum likelihood estimate of the changepoint parameter τ is the value that maximizes $M(j)$.

Our proof of Theorem 4.5 in the supplemental article [Bickel et al. (2010)] shows that, in the case where there is one true change in mean at τ, the increase in the variance estimated by block subsampling with block length L, given no segmentation [i.e., $\mathbf{t}^{(n)} = \{0, n\}$] over the variance estimated by Algorithm 4.3 conditioned on a change-point at τ, is $LM(\tau) + o_p(1)$. Subsampling conditioned on any segmentation $t \neq \tau$ would inflate the variance estimate. Hence, segmenting at $\hat{\tau} = \arg\max_j M(j)$ is optimal in the sense that $\hat{\tau}$ is the changepoint estimate that minimizes the asymptotic error of the block subsampling variance estimate. This fact does not require the assumption of independence observations, and is true for any second order stationary sequence. Thus, if we knew that there were only one changepoint, and if the goal of the segmentation is to obtain the best stratified variance estimate, then the best place to segment is $\hat{\tau}$. The block subsampling variance estimate, given the segmentation $\{0, t, n\}$, would be

$$\begin{aligned} V(t) = {} & \left(\frac{t}{n^2}\right) \sum_{i=1}^{t-tL/n} (\bar{X}_{i:i+tL/n} - \bar{X}_{1:t})^2 \\ & + \left(\frac{n-t}{n^2}\right) \sum_{i=t+1}^{n-(n-t)L/n} (\bar{X}_{i:i+(n-t)L/n} - \bar{X}_{t+1:n})^2. \end{aligned} \tag{4.8}$$

The dyadic segmentation procedure recursively applies the above logic, as described below.

ALGORITHM 4.7. Fix minimum region length $0 < L_s < n$ and threshold $b > 0$. Initialize $\mathbf{t} = \{t_0 = 0, t_1 = n\}$. Repeat:

1. For $i = 1, \ldots, |\mathbf{t}| - 1$, let $M^{(i)}(j)$ and $V^{(i)}(j)$ be respectively the processes (4.7) and (4.8) computed on the subsequence $X_{t_{i-1}+1}, \ldots, X_{t_i}$. If $t_i - t_{i-1} > 2L_s$, then let $t_i' = \arg\max_{t_{i-1}+L_s<j<t_i-L_s} M^{(i)}(j)$, $B_i = M^{(i)}(t_i')$ and $V_i = V^{(i)}(t_i')$. Otherwise, let $B_i = 0$, $V_i = \infty$.

2. Let $\lambda_i = L(t_i - t_{i-1})/n$, and
$$J_i = 1\left(\frac{(t_i - t_{i-1})B_i}{\sqrt{V_i\hat{\lambda}_i}} > b\right).$$
If $\prod_i J_i = 0$, stop and return $\mathbf{t}$.
3. Let $i^* = \arg\max_i B_i$, and $t^{\text{new}} = t'_{i^*}$.
4. Let $\mathbf{t} = \mathbf{t} \cup t^{\text{new}}$, reordered so that t_i is monotonically increasing in i.

Each step of the recursion in Algorithm 4.7 proceeds as follows: In step 1, $M^{(i)}(j)$, the generalized likelihood ratio process, and $V^{(i)}(j)$, the block subsampling variance process, are computed for each segment $[t_{i-1}+1, t_i]$ of the current segmentation. For each segment i, B_i is the maximum squared difference in mean for segment i, t'_i is the changepoint estimate that achieves this maximum, and $\hat{\lambda}_i V_i$ is the estimate of variance given a changepoint at t'_i. In computing B_i and V_i we do not allow break points that create a region with length less than L_s. In step 2, we normalize the statistic $(t_i - t_{i-1})B_i$ by the estimated standard deviation $\sqrt{\hat{\lambda}_i V_i}$. If this normalized statistic is below the threshold b for every subsegment, then the recursion stops and returns the current segmentation. Otherwise, in step 3, the optimal changepoint over all regions $t^{(\text{new})}$ is chosen to be the cut that maximizes the decrease in error of the block subsampling variance estimate. In step 4, this new changepoint $t^{(\text{new})}$ is added to the current segmentation $\mathbf{t}$.

The computation of V_i in step 2 requires an appropriate choice $L = L_b$ of the block subsampling sample size. If the correct segmentation is known, then the choice of L_b is easier, as described in Section 4.6. When the segmentation is not known, but a ball park value of L_b is available, then a segmentation can be computed using the ballpark value. The segmentation can then be used to obtain a better choice of L_b. If a ball park value of L_b is not available, then the normalization by V_i can be omitted, in which case the parameter b in step 3 should be set to 0. This would be equivalent to stopping the segmentation only when the next optimal cut will violate the minimum region length L_s. In the examples of Section 5.1 we set $b = 0$, thus decoupling the choice of L_s from that of L_b.

Two more parameters required by Algorithm 4.7 are L_s and b. The choice for L_s is discussed in Section 4.5. The choice of b can be guided by the fact that, under the null hypothesis, if L were chosen appropriately, then $(t_i - t_{i-1})M^{(i)}(j)/[V^{(i)}(j)\hat{\lambda}_i]^{1/2}$ is a pivot with approximate distribution χ^2_1. Asymptotic approximations for the family-wise error rate have been derived in the case of independent sequences [James et al. (1987)]. In the case of dependent sequences a Bonferroni adjustment can be applied to adjust for multiple testing. We also used the formulas given in James et al. (1987) to get a crude cutoff, which seems to work in practice.

Algorithm 4.7 belongs to the class of dyadic segmentation algorithms for detection of changepoints, the consistency of which were studied by Vostrikova (1981). These algorithms are greedy procedures that avoid the search over all possible segmentations. They have been applied successfully to various settings in biology, including segmentation of GC content [Li et al. (2002)] and the analysis of DNA copy number data [Olshen et al. (2004)].

The consistency of Algorithm 4.3 requires conditions A9–A11 to be satisfied by the estimated segmentation. The key condition is A11 which defines a consistency criterion on the segmentation. Consistency of dyadic segmentation has been proved in Vostrikova (1981) for sequences that satisfy the following conditions:

1. Let $\epsilon_t = X_t - E[X_t]$, then $\|\epsilon_t\|^2$ is a submartingale and $E\|\epsilon_t\|^2 < ct^\beta$, $c > 0$, $\beta < 2$.
2. The number of regions is fixed and the region sizes are of order n, that is,

$$\tau_n = (nr_1, \ldots, nr_U), \qquad 0 < r_1 < \cdots < r_U.$$

It is easy to verify that condition 1 is satisfied by the piecewise stationary model due to the mixing condition A2. Condition 2 is more stringent than our assumptions A3 and A4, under which U_n is allowed to increase with n. The consistency of dyadic segmentation for the case of $U_n \to \infty$ has been explored in Venkatraman (1992), who gave asymptotic conditions on the rejection threshold and on the sizes of the regions to ensure consistency under the assumption of an independent Gaussian sequence. However, these conditions are hard to verify in practice, and for many applications in genomics the more stringent condition of Vostrikova (1981) is sufficient. Previous studies on segmenting the genome based on features such as the GC content [Fu and Curnow (1990); Li et al. (2002)] have used this finite regions assumption to achieve reasonable results.

The dyadic segmentation procedure uses information from the entire sequence to call the first change, and then recursively uses all of the information from each subsegment to call each successive change in that segment. An alternative is to use pseudo-sequential procedures, which are sequential (online) schemes that have been adapted for changepoint detection when the entire sequence of a fixed length is completely observed. The basic idea of this class of methods is to do a directional scan starting at one end of the sequence. Every time a changepoint is called, the observations prior to the changepoint are ignored and the process starts over to look for the next change after the previously detected changepoint. Specifically, let $\hat{\tau}_0 = 0$ and, given $\hat{\tau}_1, \ldots, \hat{\tau}_k$,

$$\hat{\tau}_{k+1} = \inf\{l > \hat{\tau}_k : S(X_{\hat{\tau}_k}, X_{\hat{\tau}_{k+1}}, \ldots, X_{\hat{\tau}_l}) > b\},$$

where S is a suitably defined changepoint statistic and b is a pre-chosen boundary. The estimates from pseudo-sequential schemes depend on the direction in which the data is scanned. Thus, while they may be suitable for, say, timeseries data, they may not be natural for segmentation of genomic data, which in most cases do not have an obvious directionality. The consistency of pseudo-sequential procedures has been studied by Venkatraman (1992), who gave conditions on $b = b_n$ and $\hat{\boldsymbol{\tau}}^{(n)}$ for consistency of $\hat{\boldsymbol{\tau}}^{(n)}$ under the setting that X_i are independent Gaussian with changing means.

4.4. *Testing the null hypothesis of no associations.* Here we extend the results in Section 4.2 to testing the null hypothesis of no association using nonlinear statistics. As we discussed in Section 1.2, the inference problem typically posed in high-throughput genomics is that of association of two features. In terms of our framework we have two 0–1 processes $\{I_k\}_{k=1,\ldots,n}$ and $\{J_k\}_{k=1,\ldots,n}$ both defined on a segment of length n of the genome. We assume that the joint process $\{I_k, J_k\}$ is piecewise stationary and mixing and want to test the hypothesis that the two point processes $\{I_k\}_{k=1,\ldots,n}$ and $\{J_k\}_{k=1,\ldots,n}$ are independent. We have studied two fairly natural test statistics in ENCODE, the "percent basepair overlap,"

$$B_n = \frac{\sum_{k=1}^{n} I_k J_k}{\sum_{k=1}^{n} I_k},$$

and the "regional overlap," R_n, which we define in Section 3, with large values of these statistics indicating dependence. The major problem we face in constructing a test is what critical values $o_{n\alpha}, r_{n\alpha}$ we should specify so that

$$P_{H_0}[B_n \geq o_{n\alpha}] \approx \alpha, \tag{4.9}$$

where H_0 is the hypothesis that the vectors $(I_1, \ldots, I_n)^T$ and $(J_1, \ldots, J_n)^T$ are independent, and the corresponding $r_{n\alpha}$ for R_n.

We aim for statistics based on B_n, R_n (respectively) which are asymptotically Gaussian with mean 0 under H_0. In general, we have to be careful about our definition of independence. If we interpret H_0 as we stated, simply as independence of the vectors $(I_1, \ldots, I_n)^T$ and $(J_1, \ldots, J_n)^T$, then

$$E_{H_0}(B_n) \approx \frac{\sum_{i=1}^{U} \sum_{k=1} n_i E_{H_0}(I_{ik}) E_{H_0}(J_{ik})}{\sum_{i=1}^{U} \sum_{k=1} n_i E_{H_0}(I_{ik})},$$

where I_{ik} and J_{ik} refer to the kth basepair in the ith segment and, hence, we have

$$E_{H_0}(B_n) \approx \frac{\sum_{i=1}^{U} \lambda_i E_{H_0}^{(i)}(I) E_{H_0}^{(i)}(J)}{\sum_{i=1}^{U} \lambda_i E_{H_0}^{(i)}(I)}. \tag{4.10}$$

The natural estimate of this expectation is then

$$\frac{1}{\bar{I}}\sum_{i=1}^{U}\lambda_i\bar{I}_i\bar{J}_i,$$

where $\lambda_i \equiv \frac{n_i}{n}$, $\bar{I}_i$ is the average of I_{ik}, $\bar{J}_i$ is the average of J_{ik}, and $\bar{I}$ is the grand average. We assume the correct segmentation.

We proceed with construction of a test statistic and estimation of the null distribution. In view of (4.10), our test statistic based on B_n is

$$T_n^O \equiv n^{1/2}(B_n - \widetilde{J}_n), \tag{4.11}$$

where

$$\widetilde{J}_n \equiv \left(\sum_{i=1}^{\hat{U}}\hat{\lambda}_i\hat{\bar{I}}_i\hat{\bar{J}}_i\right)\Big/\frac{1}{n}\sum_{k=1}^{n}\hat{I}_k, \tag{4.12}$$

$$\text{where } \hat{\lambda}_i = \lambda_i(\hat{\mathbf{t}}), \hat{\bar{I}}_i = n_i^{-1}(\hat{\mathbf{t}})\textstyle\sum_{k=\hat{t}_{i-1}+1}^{\hat{t}_i} I_k$$

with $\hat{\bar{J}}_i$ similarly defined. Here is the algorithm based on this statistic.

ALGORITHM 4.8. In order to estimate the null distribution, we do the following:

1. Pick at random without replacement two starting points, K_1 and K_2, of blocks of length L from $\{1,\dots,n-L\}$.
2. Let $(I_{K_1+1},\dots,I_{K_1+L})^T$ and $(J_{K_1+1},\dots,J_{K_1+L})^T$, $(I_{K_2+1},\dots,I_{K_2+L})^T$ and $(J_{K_2+1},\dots,J_{K_2+L})^T$ be the two sets of two feature indicators.
3. Form

$$\overline{IJ}_{nL}^{*1} \equiv \frac{1}{L}\sum_{l=1}^{L} I_{K_1+l}J_{K_2+l},$$

$$\bar{I}_{nL}^{*1} \equiv \frac{1}{L}\sum_{l=1}^{L} I_{K_1+l},$$

$$\overline{IJ}_{nL}^{*2} \equiv \frac{1}{L}\sum_{l=1}^{L} I_{K_2+l}J_{K_1+l}$$

and define $\bar{I}_{nL}^{*2}$, $\bar{J}_{nL}^{*1}$, $\bar{J}_{nL}^{*2}$ analogously. Let

$$F_{nL}^* \equiv \frac{1}{2}\left(\frac{\overline{IJ}_{nL}^{*1}}{\bar{I}_{nL}^{*1}} + \frac{\overline{IJ}_{nL}^{*2}}{\bar{I}_{nL}^{*2}}\right),$$

$$T_{nL}^* \equiv F_{nL}^* - \bar{J}_{nL}^*,$$

where

$$\bar{J}^*_{nL} = \frac{1}{2}(\bar{J}^{*1}_{nL} + \bar{J}^{*2}_{nL})$$

and $\bar{I}^*_{nL}$ is defined analogously. Let F^*_{nLb}, $\overline{IJ}^{*1}_{nLb}$, etc., be obtained by choosing (K_{1b}, K_{2b}), $b = 1, \ldots, B$, independently as usual.

4. We use the following $c_{nL\alpha}$ as a critical value for B_n at level α,

$$c_{nL\alpha} = \bar{J}_n + \left(\frac{2L}{n}\right)^{1/2} T^*_{nL(B(1-\alpha))},$$

where $T^*_{nL(1)} \le \cdots \le T^*_{nL(B)}$ are the ordered T^*_{nLb} and $[\cdot]$ denotes integer part and $\bar{J}_n = \frac{1}{n}\sum_{k=1}^n J_k$.

5. If the sequence is piecewise stationary with estimated segments $j = 1, \ldots, \hat{U}_n$ as in Section 4.3, we draw independently B sets of starting points, $K^{(j)}_{11}, \ldots, K^{(j)}_{1B}$ and $K^{(j)}_{21}, \ldots, K^{(j)}_{2B}$, of blocks of length $\hat{\lambda}_j L$ from each segment $i = 1, \ldots, j$ when each pair is drawn at random without replacement. Here $\sum_{i=1}^{\hat{U}} \hat{\lambda}_i = 1$ and $\hat{\lambda}_i$ is proportional to the length of estimated segment i. Then piece T^*_{nLb} together as follows. Let

$$\overline{IJ}^{*1i}_{nLb} = \frac{1}{L\hat{\lambda}_i}\sum_{l=1}^{\hat{\lambda}_i} I_{iK_{1b}+l} J_{iK_{2b}+l},$$

$$\bar{I}^{*1i}_{nLb} = \frac{1}{L\hat{\lambda}_i}\sum_{l=1}^{L} I_{iK_{1b}+l},$$

$$\text{etc.},$$

$$\bar{F}^*_{nLb} = \sum_{i=1}^{\hat{U}} \hat{\lambda}_i \left(\frac{\overline{IJ}^{*1i}_{nLb}}{\bar{I}^{*1i}_{nLb}} + \frac{\overline{IJ}^{*2i}_{nLb}}{\bar{I}^{*2i}_{nLb}}\right).$$

Then,

$$T^*_{nLb} = F^*_{nLb} - \widetilde{J}^*_{nLb},$$

where

$$\widetilde{J}^*_{nLb} = \frac{\sum_{i=1}^{\hat{U}} (\bar{I}^{*i}_{nLb})(\bar{J}^{*i}_{nLb})\hat{\lambda}_i}{\sum_{i=1}^{\hat{U}} (\bar{I}^{*i}_{nLb})\hat{\lambda}_i}$$

with $\bar{I}^{*i}_{nLb} = \bar{I}^{*1i}_{nLb} + \bar{I}^{*2i}_{nLb}$. The critical value is

$$\widetilde{J}_n + \left(\frac{2L}{n}\right)^{1/2} T^*_{nL(B(1-\alpha))},$$

as before.

We can apply this principle more generally to statistics which are functions of sums of products of I's and J's evaluated at the same positions.

The proof of the following theorem is given in the supplemental article [Bickel et al. (2010)].

THEOREM 4.9. *If $\mathcal{L}_0$, P_0 denote distributions under the hypothesis of independence and* A1–A11 *hold, then*

1. $\mathcal{L}_0(T_n^O) \Longrightarrow \mathcal{N}(0, \sigma_0^2)$
2. *With probability tending to 1,*
$$\mathcal{L}_0^*(T_{n,L}^{O*}) \Longrightarrow \mathcal{N}(0, \sigma_0^2).$$
3. $P_0[T_n^O \geq (\frac{2L}{n})^{1/2} \hat{q}_{1-\alpha}^0] \to \alpha$ *where* $\hat{q}_{1-\alpha}^0$ *is the* $[(1-\alpha)B]$*th of* T_{nLb}^{O*}, $1 \leq b \leq B$.

In practice, this definition of independence makes our statistic in effect reflect *conditional independence* of I_k and J_k given the segment to which the kth base belongs. This can be unsatisfactory in practice, for instance, when the features are concentrated in small segments such that large, sparse segments swamp the inference.

We define *independence irrespective of segment identity* as saying that the average over all permutations of the segments of the joint distribution of the point process features are independent. Formally, if $(P_1, \ldots, P_U)$, $(Q_1, \ldots, Q_U)$ denote the marginal distributions of $\{\{I_{ik} : k = 1, \ldots, n_i\} : i = 1, \ldots, U\}$ and $\{\{J_{ik} : k = 1, \ldots, n_i\} : i = 1, \ldots, U\}$, and $(R_1, \ldots, R_n)$ correspond to the joint distribution of $\{(I_{ik}, J_{ik}) : 1 \leq k \leq n\}$, then let $(\bar{P}_1, \ldots, \bar{P}_U) = \frac{1}{U!} \sum (P_{\pi 1}, \ldots, P_{\pi U})$ where π ranges over all permutations of $1, \ldots, U$. Define $(\hat{Q}_1, \ldots, \hat{Q}_U)$ and $(\hat{R}_1, \ldots, \hat{R}_U)$ similarly. Then, our hypothesis is

$$H_1 : \hat{R} = \hat{P} \times \hat{Q}. \tag{4.13}$$

This is simply saying that independence is not conditional on relative genomic position of segments.

It is easy to see that we should now define

$$T_n^{\widetilde{O}} = n^{1/2}(B_n - \hat{J}_n), \tag{4.14}$$

where $\hat{J}_n = \frac{1}{n} \sum_{i=1}^n J_i$.

The reason for this is that

$$E_{\hat{R}}(B_n) \approx \frac{E_{\hat{R}}(\frac{1}{n} \sum_{i=1}^n I_i J_j)}{E_{\hat{R}}(\hat{I})}. \tag{4.15}$$

Under H_1,

$$E_{\hat{R}}\left(\frac{1}{n} \sum_{i=1}^n I_i J_j\right) = E_{\hat{P}}(\hat{I}) E_{\hat{Q}}(\hat{J})$$

and

$$E_{\hat{R}}(I) = E_{\hat{P}}(\hat{I}),$$

so that the statistic simplifies to the $U = 1$ form, as above.

It is clear that the conclusion of (4.9) continues to hold when applied to $T_n^{\widetilde{O}}$. Note that the form of the bootstrap is unchanged, since $T_n^{\widetilde{O}}$ is invariant under permutation of the segments.

We now turn to R_n as defined in Section 3. We assume that $V_i : i = 1, \ldots, K$ are strongly mixing and stationary. If we assume H_0, we have no closed form for $E_{H_0}(\frac{1}{W}\sum_{k=1}^{K} V_k)$ by which to center R_n. To estimate this quantity, we apply a version of the double bootstrap [Beran (1988); Hall (1992); Letson and McCullough (1998)].

Consider $\frac{1}{n}\sum_{k=1}^{K} V_k$ under H_1. We draw B_1 pairs of large blocks of length mL, and we compute the % false region overlap, call it $R_b^*, b = 1, \ldots, B$, in each pair of "large" blocks, where mL is still negligible compared to segment size, but $m \to \infty$. Define

$$\hat{E}_{H_1}(R_n) = \frac{1}{2B}\sum_{b=1}^{B_1} R_b^* \tag{4.16}$$

and

$$\widetilde{T}_n^{(R)} = n^{1/2}(R_n - \hat{E}_{H_1}(R_n)). \tag{4.17}$$

Note that we again want to consider independence irrespective of segment identity, so that R_b^* above are computed without any segmentation beyond the natural segmentation, for example, chromosomes. Now compute the empirical distribution of $\widetilde{T}_n^{(R)}$ using the size L segmented block subsampling and proceed as usual. We can define $\widetilde{T}_n^{(R)}$ corresponding to H_0 in the same way, though we now have to cut up our mL blocks in proportion to segment sizes to center. We do not pursue this since the H_1 hypothesis gives stable results while H_0 does not.

We have not proved a result justifying the use of the double bootstrap in this way, but simulations suggest that it behaves as expected; see Section 5.3.

4.5. *Choice of segment size L_s.* Two tuning parameters appear in our procedure in addition to b appearing in the segmentation scheme. L_s is the smallest allowed size of a "stationary" piece after segmentation. It essentially determines the scale of the segmentation, which we view as an application context dependent quantity that users need to control. The reason is that stationarity is a matter of scale. To put it concretely, consider the situations where I_k, $k = 1, \ldots, n$, are simply the base pair nucleotides A, C, G, T and consider the scale of a large gene of length n. Then, it seems natural that the

exons and introns correspond to consecutive stationary regimes. However, suppose we now move our scale to a gene rich genomic region of length N. Now, it is the genes themselves and the intergenic regions which correspond to an at least initial segmentation.

This dependence of segmentation on scale has a natural intuitive consequence. Consider a statistic such as base pair overlap of two features. As one increases the region size n in which one wishes to declare significant overlap, the standard deviation of the statistic, which is $O(n^{-1/2})$, decreases, and p-values decrease. However, if, as one would expect, the region over which n increases becomes homogeneous on a larger scale, coarser segmentation would then be called for. This, as we have noted, necessarily increases the standard deviation of the statistic, and from that point of view significance becomes more difficult to achieve.

Put another way, it is not impossible to think of the whole genome itself as being stationary on a large scale, but that we can hierarchically segment the genome in many ways so that each large subsegment is stationary, but the segments are not identically distributed, even where they are of equal length. For instance, a natural initial segmentation is to chromosomes.

Finally, we argue in mathematical terms going the other way from inhomogeneity to homogeneity. Start with a sequence of independent (say) Bernoulli variables $X_1, X_2, \ldots, X_n$, with X_k being Bernoulli(p_k). If the p_k are arbitrary, the only segmentation we can perform is the useless trivial one, where each X_k is its own segment. But, now we tell ourselves that p_k, $1 \le k \le n/2$, are drawn i.i.d. from $U(0, 1/2)$ and for $n/2 + 1 \le k \le n$ from $U(1/2, 1)$, we suddenly just have two segments to consider.

Thus, L_s in our view needs to be treated as the smallest scale on which homogeneity is expected. Note that these considerations are not limited to testing. They also govern confidence intervals, as discussed in Section 4.2.3.

4.6. *Choice of L_b, the subsample size.* We believe that the best way to choose L_b, after segmentation has been estimated, is so that the resulting subsampling distribution of the statistics is as stable as possible and L_s is large but $\ll n$. We also formally consider Gaussianity of the distribution and, if possible, maximizing that feature as well. This does not necessarily mean segment more—since A10 and A11 may then fail. We advocate but do not analyze further the following proposal put forward in m-out-of-n subsampling by Bickel, Götze and van Zwet (1997) and analyzed in detail by Götze and Rackauskas (2001) and Bickel and Sakov (2008):

1. Let $\bar{X}_n^*(L)$ be the statistic computed from the sample drawn with blocks of length L. Compute the block subsampling distribution $\mathcal{L}_{L_v}$ for the statistic

$$\sqrt{L_v}(\bar{X}_n^*(L_v) - \bar{X}_n)$$

and $L_v = \rho^v n$, where $\rho < 1$ and $v = 1, 2, \ldots, V$. Note that these L_v provide candidate choices of the subsample size L_b.
2. Compute a "distance" $d^*(v)$ between $\mathcal{L}_{L_v}$ and $\mathcal{L}_{L_{v-1}}$.
3. Choose $L_b = L_{v_0^*}$, where $v_0^* = \arg\min d^*(v)$.

In practice, we use for $d^*(v)$ the pseudometric

$$\left| \sqrt{\frac{L_{v-1}}{L_v}} IQR(\mathcal{L}_{L_v}) - IQR(\mathcal{L}_{L_{v-1}}) \right|,$$

where $IQR(\mathcal{L})$ is the interquartile range of $\mathcal{L}$.

In continuing work with Götze and van Zwet, we are in the process of trying to show that, under mild conditions, as $n \to \infty$ we have $L_b \to \infty$, $L_b/n \to 0$. More significantly, we expect that in a fashion analogous to Götze and Rackauskas (2001) and Bickel and Sakov (2008), under restrictive conditions and for suitable choice of distance, L_b yields an estimate which is as good as possible in the following sense: If $\mathcal{L}_n$ is the actual distribution of $\sqrt{n}(\bar{X}_n - \mu)$, $d(m)$ is the distance between $\mathcal{L}_n$ and $\mathcal{L}_{L_v}$, and $v_0 = \arg\min_v d(v)$, then

$$\frac{d(v_0^*)}{d(v_0)} \to_p c.$$

Thus, $L_{v_0^*} = \rho^{v_0^*} n$ yields performance of the same order as $\rho^{v_0} n$.

5. Simulation and data studies.

5.1. *Simulation study I.* In this section we perform a simple simulation study to demonstrate the power of our block-subsampling method in the situation where features are naturally clustered. We simulated a binary sequence $x_1, \ldots, x_n$ with $n = 10{,}000$ by the following Markovian model:

$$\begin{aligned} P(x_1 = 1) &= \frac{p_0}{2}, \\ P(x_k = 1) &= \frac{p_0}{2} + (1 - p_0)\frac{\sum_{j=k-w}^{k-1} x_j}{w} \qquad \text{for } k = 2, \ldots, n, \end{aligned} \tag{5.1}$$

where w is the order of the Markov model or, intuitively, the size of the dependency window, and p_0 indicates the level of dependency. Smaller p_0 gives stronger dependence between neighboring positions. We define the following two types of features at position k in the sequence:

- Feature I: the occurrence of sequence 11,100 starting at position k.
- Feature II: the occurrence of more than six 1's in the next 10 consecutive positions including the current position k.

From model (5.1), the feature II will occur in clusters in the sequence. The overlap between the two types of features can be measured by the statistic

$$S = \frac{\sum_{k=1}^{n} I_k J_k}{\sum_{k=1}^{n} I_k}$$

with I_k, J_k being binary and indicating the occurrences of sites of types I and II, respectively.

Figure 1 shows the distribution of S estimated through different ways:

- The true distribution is the empirical distribution of estimated S from 10,000 random sequences generated under model (5.1).
- The Ordinary Bootstrap distribution is derived by performing a base-by-base uniform sampling of the sequence $x_1, \ldots, x_n$ to construct 10,000 sequences of length n.
- The Feature Randomization distribution is derived by keeping features of type I fixed and randomizing uniformly the start positions of the features of type II to construct 10,000 sequences of length n.
- The block subsampling distribution is derived by drawing independent samples of blocks of length $L = 40$ and stringing the blocks together to construct 10,000 sequences of length n.

From Figure 1, we see that block subsampling produces more reliable estimates of the variance of S compared to the naive methods: ordinary bootstrapping and feature randomization. Both naive methods ignore the dependence between positions and thus fail to take into account the natural clumps of the feature II. This is the key reason for the poor performance of the two naive methods.

5.2. *Simulation study IIa.* Our second simulation study examines the case where the sequence is generated from a piecewise stationary model where there is more than one homogeneous region. As before, we consider the problem of estimating the percentage of base pair overlap between two features, and compare the performance of four strategies:

1. feature randomization,
2. naive block subsampling from unsegmented sequence,
3. block subsampling from sequence segmented using the true changepoints, and
4. block subsampling from sequence segmented using the changepoints estimated by the dyadic segmentation method we described in Section 4.3.

In our simulation model, we generate X_t, Y_t independently from a Neyman–Scott process characterized as follows:

1. Cluster centers occur along the sequence according to a Poisson process of rate λ_i in region i.

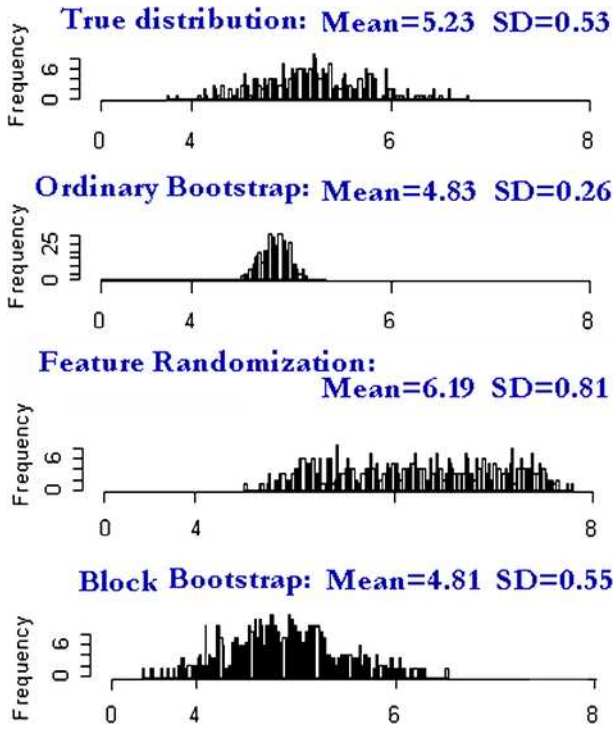

FIG. 1. *Comparison of different subsampling schemes.*

2. The number of features in each cluster follows Poisson distribution with mean α.
3. The start of features are located at a geometric distance (mean μ) from the cluster center.
4. The features are generated with length that is geometric with mean β.
5. Overlap between features generated using steps 1–4 are ignored.

For simplicity, we let there be only 2 homogeneous regions, each of length $T = 10{,}000$. Consider the setting where the parameters for the two regions have the following values: $(\lambda_1, \alpha_1, \mu_1, \beta_1) = (0.01, 10, 10, 5)$ and $(\lambda_2, \alpha_2, \mu_2, \beta_2) = (0.02, 10, 10, 5)$. Figure 2 shows a simulated example, where features A and B are plotted as well as their overlap. Figure 2 also shows the cumulative sum and the segmentation. Figure 3 shows respectively the histograms of the estimated distribution of the overlap statistic $\bar{X}^*$ centered and scaled. It is clear that the feature randomization underestimates the standard deviation, whereas naive block subsampling without segmentation gives a mixture distribution with long tails. Strategy 3, which subsamples assuming the true changepoint at τ is known, gives the correct distribution as expected. Strategy 4, which uses the estimated changepoint, reassurringly gives a very

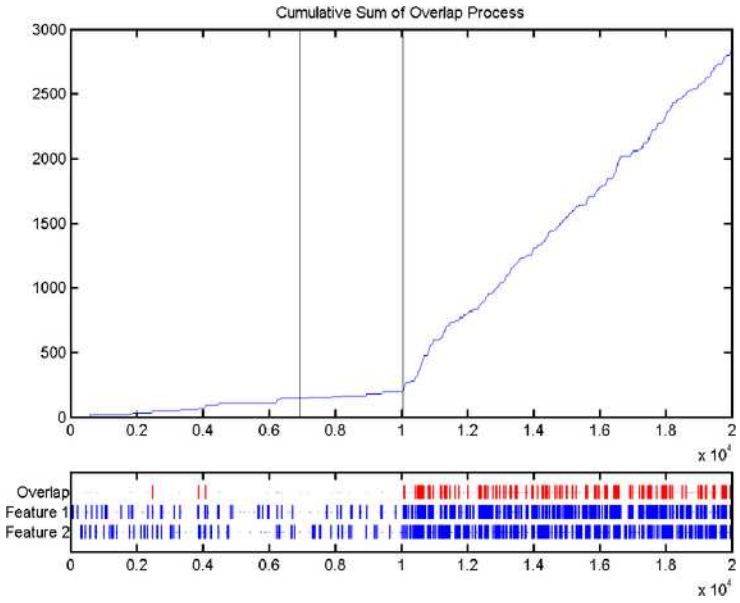

FIG. 2. *Example of one instance from simulation model 2. Top plot shows cumulative sum and estimated segmentation.*

similar distribution to strategy 3. Table 1 gives the standard deviation estimates.

5.3. *Simulation study IIb.* We utilized the Neyman–Scott process described in simulation study IIa to study the consistency of the double bootstrap method described in Section 4.4 for estimating the distribution of R_n. We consider the simple case where there is one homogeneous region. We utilized a larger region and a parameterization of the process that results in more and longer feature instances than we consider in the study above. The parameters are $T = 5$ Mb and $(\lambda_1, \alpha_1, \mu_1, \beta_1) = (\lambda_2, \alpha_2, \mu_2, \beta_2) = (0.05, 10, 100, 75)$. This yields a pair of feature-sets with around 5000 instances, where each feature-set covers around 17% of the 5 Mb region. We simulated 20,000 pairs of feature-sets from this process, and found that the mean of region-overlap between pairs, R_n, was 0.293, and the standard error was 0.0072. We subsampled 1000 sets of 10,000 draws from this distribution, each of which yielded the mean above (to 3 significant digits), and the standard errors ranged from 0.0071 to 0.0073, which corresponds almost exactly to the theoretical 95% confidence interval for the standard error of the standard error of a Gaussian with standard deviation 0.0072 after 10,000 draws. Not surprisingly, the distribution of R_n was Gaussian, as indicated by the Lilliefors and the Shapiro–Wilk test, which did not reject the hypothesis

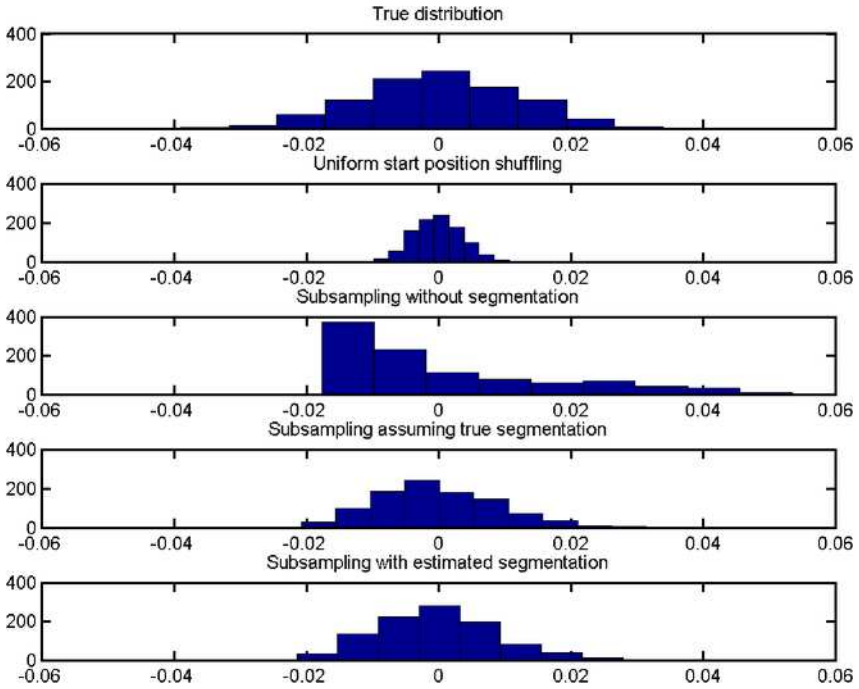

FIG. 3. *Comparison of different subsampling schemes.*

of Gaussianity at a significance level 0.05 with the full sample of 20,000 observations.

In order to test the capacity of segmented subsampling with a version of the double block bootstrap to discover this distribution based on only a single pair of observations, we selected the most extreme pair found during simulation, for which R_n was 0.321, corresponding to a z-score of 3.87. Since the number of feature instances is itself a random quantity, the job of block subsampling is particularly difficult: when R_n is far to the right of expectation, the feature-sets tend to contain more feature instances than those closer to the center. The pair we chose was no exception. The results are given in Figure 4. Hence, it is not surprising that our subsampling procedure tends to over-estimate the mean. The Lilliefors test fails to reject the Gaussianity of any of the resulting distributions with sample sizes up to 1000 at a significance level of 0.05, but does reject it for several of the smaller block-sizes when the sample size is pushed up to 10,000. The Shapiro–Wilk test, however, detects departures from Gaussianity for many of the distributions at a significance level of 0.05 for samples larger than 500. This is because R_n is predicated on relatively small counts of feature-instance overlaps and, hence, the distributions tend to have heavy tails.

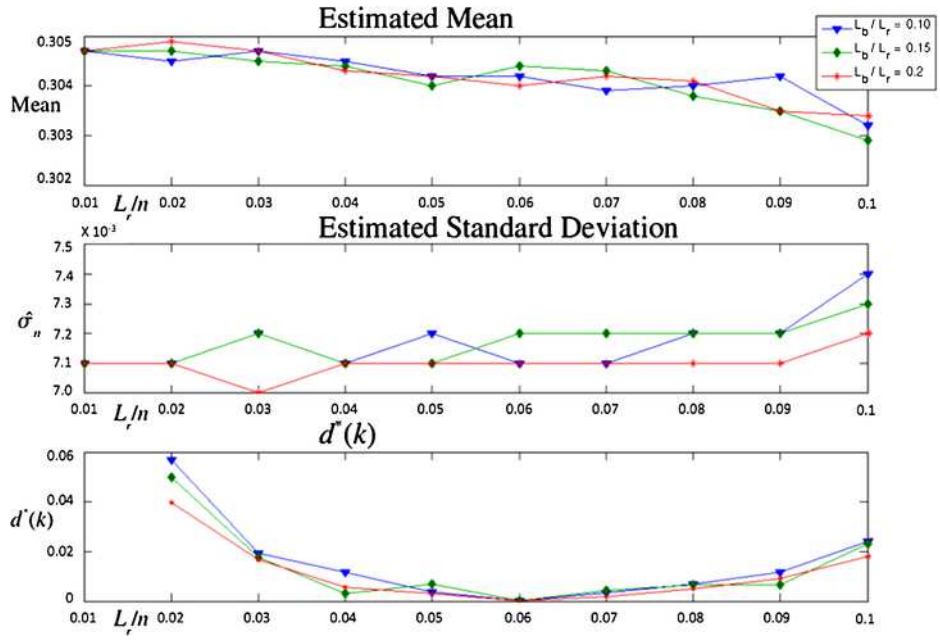

FIG. 4. *Comparison of block subsampling distributions.*

We note that the global minimum of the Inter-Quantile (IQ) statistic was found at $L_b/L_r = 0.15$ and $L_r/n = 0.06$. That is, 0.9% of the 5 Mb region, or 45 Kb, were included in each block sample. This block sample size is certainly sufficient to capture multiple feature-clusters, since the parameterized Neyman–Scott process above yields an average inter-cluster distance of about 1 Kb.

To corroborate our hypothesis that the mean was overestimated because the feature-sets we chose were more dense than most, we applied our method with learned parametrization, $L_b/L_r = 0.15$ and $L_r/n = 0.06$, for a pair of

TABLE 1
Estimates of standard error by four sampling strategies in simulation study IIa

Method	Standard error estimate	Fold change from true value
True value	1.2e−002	—
Uniform shuffle	3.6e−003	0.3
Subsample, no segmentation	1.7e−002	1.4
Subsample, true segmentation	1.1e−002	0.91
Subsample, estimated segmentation	1.0e−002	0.83

feature-sets with $R_n = 0.293$, the population average. Indeed, the mean was estimated, after 10,000 samples, to be 0.293, and $\hat{\sigma}_n$ was 0.0072.

Although the purpose of this simulation was merely to check the consistency of our version of the double block bootstrap for data not unlike actual genomic data, for example, ChIP-seq "broad-peaks," we decided to also check the performance of feature-start site shuffling for the same pair of feature-sets used above. In the case of B_n, the basepair overlap statistic, feature start-site shuffling correctly estimates the mean, but can (in the stationary case) radically underestimate the standard deviation. The same is not true in the case of R_n. Start-site shuffling is not assured (under our model) to provide an unbiased estimate of the mean or the standard deviation. We drew 10,000 samples from the distribution under shuffling, and found the mean to be 0.337, and the standard deviation to be 0.0070, which indicates that the pair of feature-sets under study in fact overlap slightly less than expected at random ($p \approx 0.011$). The fact that this conclusion is actually in the wrong direction in this relatively easy, stationary example should make us skeptical of studies that rely upon start-site shuffling to draw conclusions about statistics that cannot be defined locally, such as R_n.

Our discussion of this simulation and the following real data examples exhibit the subtleties inherent in our approach. Subtleties appear whenever inference follows regularization.

5.4. *Association of noncoding ENCODE annotations and constrained sequences.* Here we present a real example of the study of association between "constrained sequences" and "nonexonic annotations" from the ENCODE project, limited to the 1.87 Mbp ENCODE Pilot Region ENm001, also known as the CFTR locus. The constrained sequences are those highly conserved between human and the 14 mammalian species studied and sequenced by the ENCODE consortium. Enrichment of evolutionary constraint at the "nonexonic annotations" sites implies that the biochemical assays employed by the ENCODE consortium are capable of identifying biologically functional elements. We tested the association of noncoding annotations and constrained elements using the base pair overlap statistic B_n in Section 4.3 using the conditional formulation. We interpret the lack of association as, given sequence composition and the distribution of each feature along the genome as observed, the assignments (by nature) of features A and B to individual bases are made independently. We derive the significance of the observed statistic under this null hypothesis following the method proposed in Section 4.3.

As we discussed, we have several issues to deal with:

(i) How do we segment? That is, what statistic(s) do we use for segmenation?

(ii) Is segmentation necessary or is the region sufficiently homogeneous?
(iii) If we segment, what L_s should we use?
(iv) Given a segmentation, what L_b is appropriate?

Here are our methods:

(a) The simplest choice for (i) and the one we followed was to segment according to both numerator and denominator in B_n: intersect partitions and enforce an L_s bound. Given our theory, this should ensure homogeneity in the mean of B_n.
(b) Although strictly speaking (ii) and (iii) can be combined, we experimented a bit to also see if the theory of Section 4.1 was borne out in practice.
(c) We did not use the V statistic and thus only had to choose L_s. Again, we experimented with $L_s = 500$ Kb to preserve as much genomic structure as possible, and $L_s = 200$ Kb to ensure we had not undersegmented.
(d) We explored a variety of values of L_b, and studied the consistency between nearby values under the interquartile statistic (IQ statistic) discussed in Section 4.6. We draw conclusions based on the value of L_b that optimizes local consistency.

To segment the data, we applied the method in Section 4.3 to both features A and B, or in the language of Section 4, I and J, and then combined the segmentation. In segmenting each feature, we experimented with minimum segment lengths L_s of 200 and 500 Kb. Before subsampling, we combined the segmentations of A and B by taking a union of the changepoints. This created regions with length less than L_s. However, the total length of these regions comprise $<0.1\%$ of the total Encode region, and were left out of the remaining analyses.

If the sequence were sufficiently homogeneous, we could forgo the initial segmentation step. Figure 5 shows an estimate of variance of B_n (with the appropriate renormalization) for a reasonable range of L_b, both before and after segmentation. Two trends are clearly evident. First, segmentation greatly reduces the estimated variance. As we discussed in Section 4.1.2, inhomogeneity of the sequence causes an inflated estimate of variance. If the data were homogeneous, segmentation should not change the variance estimate. Thus, the fact that the estimated variances drop after segmentation for such a large range of L_b's suggests that the data are inhomogeneous. Second, and more importantly, the estimated variance of B_n increases sharply with increasing L_b in the unsegmented data. This is evidence of inhomogeneity in the mean of B_n across this ENCODE region: underlying shifts in mean, if ignored, can be mistaken for spurious long range autocorrelation, which also implicitly runs against our assumption. In either case, as Theorem 4.2 suggests, we would be overly conservative. Thus, a preliminary exploration

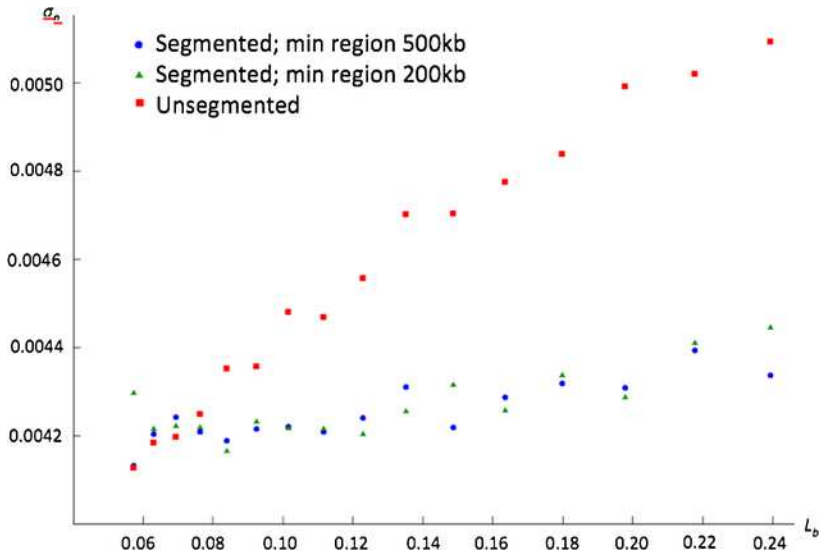

FIG. 5. *Estimated σ_n as a function of L_b for 10,000 samples.*

of the data convinces us that this ENCODE region is inhomogeneous in I and/or J and segmentation is necessary.

We found that 200 and 500 Kb gave 5 and 3 segments respectively. Figure 6 gives the results for 500 Kb. What is fairly surprising, but reassuring, is that over the whole broad range of L_b considered, the estimated SD of the statistic under the null was essentially flat after segmentation. Flat here means that variability was within a Monte Carlo SD for the 10,000 replications we used. We would expect longer values of L_b to include, in our estimate of σ, additional covariance between distant genomic positions captured by the extended block-length. The fact that this, by and large, does not appear to be happening is consistent with our hypothesis that the relevant mixing distance is indeed quite small compared to the size of approximately stationary regimes.

We found that there is still moderate deviation from Gaussianity in both the segmented and unsegmented case for $0.05 < L_b < 0.25$, both in the tails, as detected by the Shapiro–Wilk test, and in the body of the distribution under the Lilliefors test. With a sample size of 100, neither test detects this departure, but at a sample size of only 500, it is detected under a number of parameterizations of L_b. As we discussed in Section 4.5, the definition of stationarity depends on the scale at which we view the genome. This suggests that our segmentation still does not take care of inhomogenity in the variance. Hence, as we have mentioned, if we use the variance for the Gaussian approximation, our results are still conservative.

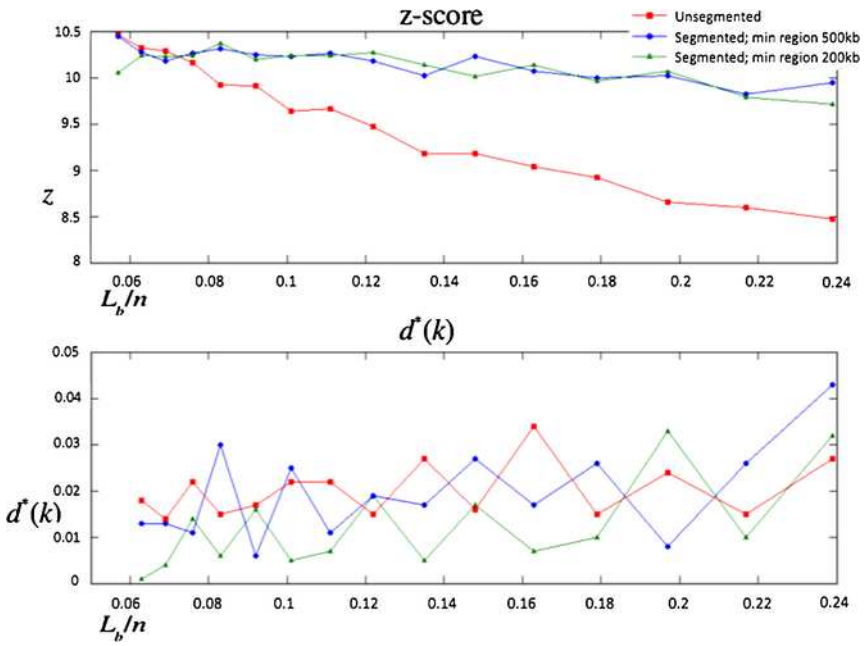

FIG. 6. *Comparison of block subsampling distributions, $\rho^\beta n$ vs. $\rho^{\beta+1}n$ under the IQR statistic. Estimates $\hat{\sigma}_n$ and resulting z-scores of B_n shown.*

The scientific conclusion of this example is that, indeed, there is strong association since the z-value is over 9 SDs. We note that the effect of segmentation on our scientific conclusion is essentially nonexistent. However, it is comforting to note that the change in (with and without segmentation) variance is in the correct direction.

5.5. *The association of copy number variation with RefSeq annotated exons in the human genome.* In this example, we reanalyze a published data set; this reanalysis leads to a different conclusion from the one made by the original paper. In 2006, Redon et al. published a set of 1445 genomic regions with observed Copy Number Variation (CNVs) across individuals. These regions consist of both deletions and insertions, and more than half of them overlap genes. In the paper, the authors reported, among other things, a paucity of overlap with RefSeq genes at a significance level of 0.05. The statistic that they used is precisely our marginal formulation of the region overlap statistic R_n, but the null distribution to which they referred it is quite different. Their null was computed by randomly permuting both genes and CNVs, and hence treats the entire genome (or at least entire chromosomes) as homogeneous, and the distances between feature-instances as exponential. Thus, if feature-instance lengths were all 1 bp, this would be

a Poisson process. As discussed in Section 5.3, under our model this procedure provides an unbiased estimate of the mean in the case of the B_n, but is unpredictable with respect to its estimate of the variance. In the case of R_n, it is unpredictable with respect to both the mean and the variance. Here, for comparison with the result of Redon et al. (2006), we examine only R_n.

Although we have attempted to replicate this portion of the Redon study, undoubtedly there are small differences between our efforts and those of Redon et al. (2006). For instance, we have masked all genomic repeats in the "Repeat Masker" track on the UCSC genome browser (genome.ucsc.edu). Redon et al. also considered patterns of repeats in their analysis, but may have utilized an at least slightly different map of genomic repeats. We find that 61.8% of the CNVs overlap RefSeq genes by at least 1 basepair. That is, we wish to assess the significance of our observed statistic $R_n = 0.618$.

The calibration of the subsampling procedure is nontrivial, especially in this application where we must consider the additional parameter L_r. Hence, in the following we provide complete detail regarding the calibration of our method for the data of Redon et al. (2006).

As before, our analysis begins with an assessment of the need for segmentation. In this case, we are dealing with whole human chromosomes, we expect that, in general, at least some segmentation is necessary. We segmented down to a minimum segment length of 10,000,000 bps (10 Mbs), letting $L_s = 10$ Mb. The mean length of these CNVs is around 250 Kb, and they are not uniformly distributed, so we are compelled not to segment down to regions much smaller than 10 Mb by our desire to capture the appropriate spatial distribution of clusters of feature-instances. To assess the sufficiency of the resulting segmentation, we examine the Gaussianity of the segmented subsampling distributions. This examination is tied to our selection of block length.

To select an inner block length, L_b, and an outer block-length, L_r, we drew 10,000 samples for each of several lengths. We chose to use a linear, rather than exponential, scale for L_r/n: we selected 10 values from 0.01 to 0.10 in increments of 0.01. We chose three values of L_b/L_r, 0.05, 0.10 and 0.20. Each of these parameterizations yields several responses, including: an estimated z-score, $d^*(k)$, and measures of Gaussianity. In Figure 7, we plot the relationship between the estimated z-score, $d^*(k)$, L_r and L_b. Regarding the Gaussianity of the resulting distributions, at a significance level of 0.01 and a sample size of 5000, neither the Shapiro–Wilk nor the Lilliefors test rejected the null hypothesis of Gaussianity for any of the 30 explored parameterizations. To supplement our biological intuition that segmentation is necessary when whole chromosomes are considered, we used the same 30 parameterizations with the unsegmented data, and performed the same tests to check the Gaussianity of the resulting distributions. Of the 30 parameterizations, 3 showed departures from Gaussianity under Lilliefors test,

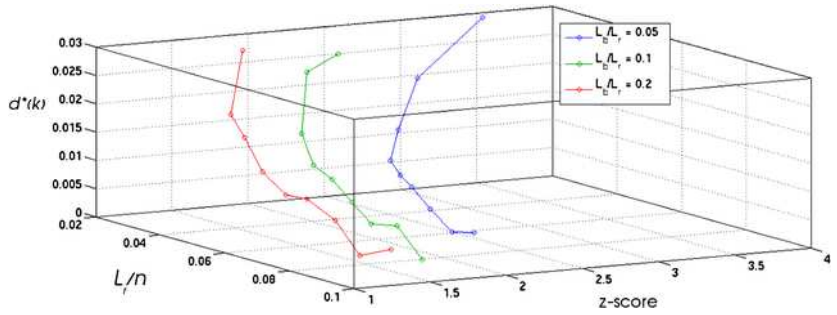

FIG. 7. *The relationship between the estimated z-score, $d^*(k)$, L_r and L_b. As L_r increases, our estimate of $\hat{\sigma}_n$ (not shown) increases, which drives the estimated z-score down. As L_r becomes too small, we lose the stability of our estimates, and $d^*(k)$ increases. For the smallest value of L_r shown here, the estimated z-score increases sharply, but the corresponding value of $d^*(k)$ indicates that this parameterization is unreliable. The ideal parameterization under $d^*(k)$ is given by $L_r/n = 0.09$ and $L_b/L_r = 0.20$.*

and 9 showed strong departures in the tails under the Shapiro–Wilks test. This indicates, as expected, that segmentation has substantially improved the Gaussianity of the sample distributions. In practice, one might attempt a finer segmentation in hopes of further reducing the (conservative) bias in $\hat{\sigma}_n$. For this example we are satisfied with the current segmentation.

The global minimum of $d^*(k)$ occurs for $L_r/n = 0.09$ and $L_b/L_r = 0.20$. This parameterization yields an estimated z-score of 1.25 and, therefore, we conclude that we cannot corroborate the result of Redon et al. (2006). Under our model it appears that CNVs are, if anything, very slightly positively associated with genes ($p \approx 0.105$). We note that a few parameterizations, as shown in Figure 7, do produce z-scores greater than 2. However, these parameterizations correspond to large values of $d^*(k)$ and, furthermore, significance is in the opposite direction reported by Redon et al. (2006). This highlights the need for carefully defined null distributions in genomic studies. We are not suggesting that the results presented necessarily invalidate the corresponding result of Redon et al. (2006), but rather we caution that scientific conclusions of this kind are predicated on how the researcher defines "at random," and that this definition should be made to reflect, as much as possible, that which is known about the actual distribution of genomic elements. We presume that authors wish, in general, to err on the side of caution, and hence do not wish to report significant association when the association can be explained simply by a conservative choice of null.

SUPPLEMENTARY MATERIAL

Some theorems in subsampling methods for genomic inference (DOI: 10.1214/10-AOAS363SUPP; .pdf). In Supplementary Material, we provide theoretical proofs to the theorems presented in the main text.

REFERENCES

ANDREWS, D. and MALLOWS, C. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **26** 99–102. MR0359122

BERAN, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83** 687–697. MR0963796

BERNARDI, G., OLOFSSON, B., FILIPSKI, J., ZERIAL, M., SALINAS, J., CUNY, G., MEUNIER-ROTIVAL, M. and RODIER, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228** 953–958.

BICKEL, P. J., BOLEY, N., BROWN, J. B., HUANG, H. and ZHANG, N. R. (2010). Supplement to "Subsampling methods for genomic inference." DOI: 10.1214/10-AOAS363SUPP.

BICKEL, P. J. and SAKOV, A. (2008). On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles. *Statist. Sinica* **18** 967–985. MR2440400

BICKEL, P. J., GOTZE, F. and VAN ZWET, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statist. Sinica* **1** 1–31. MR1441142

BIRNEY, E. ET AL. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447** 799–816.

BLAKESLEY, R. W. ET AL. (2004). An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res.* **14** 2235–2244.

BRAUN, J. and MULLER, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statist. Sci.* **13** 142–162.

CARTER, N. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genet.* **39** S16–S21.

CHURCHILL, G. A. (1989). Stochastic models for heterogeneous genome sequences. *Bull. Math. Biol.* **51** 79–94. MR0978904

CHURCHILL, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Comput. Chem.* **16** 107–115.

DAS, D., BANERJEE, N. and ZHANG, M. Q. (2004). Interacting models of cooperative gene regulation. *Proc. Natl. Acad. Sci. USA* **101** 16234–16239.

DEDECKER, J., DOUKHAN, P., LANG, G., LEON R., J. R., LOUHICHI, S. and PRIEUR, C. (2007). *Weak Dependence: With Examples and Applications. Lecture Notes in Statist.* **190**. Springer, New York. MR2338725

EFRON, B. (1981). Nonparametric standard errors and confidence intervals. With discussion and a reply by the author. *Canad. J. Statist.* **9** 139–172. MR0640014

FICKETT, J. W., TORNEY, D. C. and WOLF, D. R. (1992). Base compositional structure of genomes. *Genomics* **13** 1056–1064.

FU, Y.-X. and CURNOW, R.-N. (1990). Maximum likelihood estimation of multiple change-points. *Biometrika* **77** 563–573. MR1087847

GOTZE, F. and RACKAUSKAS, A. (2001). Adaptive choice of bootstrap sample sizes. In *State of the Art in Probability and Statistics (Leiden, 1999)* 286–309. *Lecture Notes Monogr. Ser.* **36**. Inst. Math. Statist., Beachwood, OH. MR1836566

GUPTA, M. and LIU, J. S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **102** 7079–7084.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237

HUANG, H., KAO, M. C., ZHOU, X., LIU, J. S. and WONG, W. H. (2004). Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *J. Comput. Biol.* **11** 1–14.

James, B., James, K. L. and Siegmund, D. (1987). Tests for a change-point. *Biometrika* **74** 71–84. MR0885920

Kato, M., Hata, N., Banerjee, N., Futcher, B. and Zhang, M. Q. (2004). Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.* **5** R56.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. MR1015147

Letson, D. and McCullough, B. D. (1998). Better confidence intervals: The double bootstrap with no pivot. *Amer. J. Agr. Econ.* **80** 552–559.

Li, W., Stolovitzky, G., Bernaola-Galván, P. and Oliver, J. L. (1998). Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Res.* **8** 916–928.

Li, W., Bernaola-Galván, P., Haghighi, F. and Grosse, I. (2002). Applications of recursive segmentation to the analysis of DNA sequences. *Comput. Chem.* **26** 491–510.

Margulies, E. H. et al. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17** 760–774.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

Politis, D. and Romano, J. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050. MR1329181

Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling*. Springer, New York. MR1707286

Redon, R. et al. (2006). Global variation in copy number in the human genome. *Nature* **444** 444–454.

Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74** 445–455. MR0909350

Venkatraman, S. (1992). Consistency results in multiple change-point problems. Ph.D. dissertation, Stanford Univ.

Vostrikova, L. J. (1981). Detecting disorder in multidimensional random process. *Sov. Math. Dokl.* **24** 55–59.

Yu, H., Yoo, A. S. and Greenwald, I. (2004). Cluster Analyzer for Transcription Sites (CATS): A C++-based program for identifying clustered transcription factor binding sites. *Bioinformatics* **20** 1198–1200.

Zhang, C., Xuan, Z., Mandel, G. and Zhang, M. Q. (2006). A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acid Res.* **34** 2238–2246.

Zhou, Q. and Wong, W. H. (2004). CisModule: De Novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* **101** 12114–12119.

P. J. Bickel
N. Boley
J. B. Brown
H. Huang
University of California at Berkeley
Berkeley, California
USA
E-mail: bickel@stat.berkeley.edu
npboley@gmail.com
benbrownofberkeley@gmail.com
hhuang@stat.berkeley.edu

N. R. Zhang
Stanford University
Stanford, California
USA
E-mail: nzhang@stanford.edu